Beyond Protocols: Why AI is Networking's Overdue Paradigm Shift

JP Vasseur, PhD

Sr Distinguished Engineer NVIDIA jvasseur@nvidia.com

July 2025

Abstract*: for the past decade, Artificial Intelligence (AI) and Machine Learning (ML) have been applied to networking in narrow, isolated ways—mostly anomaly detection, traffic forecasting, failure prediction—but the impact remains surprisingly and unfortunately too marginal. Core operations like troubleshooting, root-cause-analysis, and network optimization still depend on outdated, manual methods. Meanwhile, Generative AI (GenAI) and large language models (LLMs) are reshaping entire industries with agentic systems, tool learning, reasoning models, and real-time decision-making. Networking is lagging behind. Today's uses—chatbots and config assistants—barely scratch the surface of what's possible. This paper argues that the biggest opportunity in networking today is not another protocol or some algorithmic optimization—but a shift to intelligent, distributed, AI-driven systems. GenAI can unlock self-healing fabrics, predictive diagnostics, and cross-layer optimization at global scale. But the gap between potential and reality is widening fast. Now is the moment to act. The future of the Internet will be built by networks that learn, reason, and adapt. Those that don't will be left behind.*

1. Introduction

The Internet has experienced unprecedented growth in scale and complexity over the past four decades. Global internet traffic is now measured in hundreds of exabytes per month, driven by a proliferation of high-bandwidth applications and cloud-based services. The number of connected devices is projected to exceed 30 billion by 2025, each contributing to the data deluge. This demand is met by an explosion in bandwidth rates, with multi-gigabit speeds becoming common, and supported by a global footprint of over 900 hyperscale data centers. This expansion has led to networks that are more heterogeneous than ever, with diverse architectures spanning edge, core, and massive data center environments interconnected by a myriad of protocols and technologies. Consequently, designing, managing, and operating these systems has become incredibly challenging, as traditional manual approaches struggle to handle the dynamic nature of modern workloads and complex failure modes.

It is therefore difficult to understand why a technology such as ML/AI has been so poorly adopted, a topic we will review in the next section.

2. Limited Adoption of ML/AI

During the past decade, despite advancements in ML/AI, their integration into networking remains limited. Factors contributing to this include the need for high-quality labeled data, which is sometimes not available (or require significant effort) in operational networks, integration challenges with legacy infrastructure but also resistance to change and skepticism from network engineers. Still, the next section highlights specific instances where ML/AI has been successfully adopted, providing context for the discussion on untapped potential.

3. Pockets of ML/AI Adoption in Networking

3.1 ML/AI Use Cases in Network Technologies

This section examines specific instances where ML/AI have been integrated into operational networking systems. The focus is on documented deployments, with contrasts to research-oriented applications where relevant.

Wi-Fi Networks: IEEE 802.11 ML/AI applications in Wi-Fi networks are deployed primarily for network management and troubleshooting in commercial products. Standardization efforts, such as the IEEE 802.11 AI/ML Topic Interest Group (TIG), have produced reports influencing standards like 802.11bn (Wi-Fi 8), but deep integration from standards at PHY/MAC layers remains in research. That being said, Wifi ML/AI have been widely deployed on premise and in the cloud allowing to deploy a variety of use cases for million of Access Points (AP). Key deployed use cases involve anomaly detection, where ML models identify deviations in radio frequency (RF) and upper layer metrics flagging potential malfunctioning devices impacting the user experience, and automated configuration, such as dynamic channel selection using optimization algorithms. For the sake of illustration, Cisco's AI Network Analytics, part of Catalyst Center Assurance, employs advanced ML techniques for establishing dynamic baselines of key performance indicators (KPIs). Cisco’s AI-based dynamic baselining is derived from rolling four-weetelemetry windows and includes metrics such as onboarding latency, signal-to-noise ratio, and application response time. It utilizes unsupervised learning to cluster normal network behaviors from anonymized telemetry data, and deviations are computed by comparing new incoming values against these percentile-based statistical models. Time-series forecasting uses exponential smoothing and regression ensembles to predict threshold crossings, with corrective suggestions integrated into the Assurance dashboard. Juniper Mist AI applies a combination of machine learning techniques for Wi-Fi assurance, based on public Juniper documentation. Reinforcement learning is employed for radio resource management, where agents dynamically optimize RF parameters such as channel selection, transmit power levels, and bandwidth allocation in real time. Mist’s RL agents operate globally in the cloud with daily learning cycles and reward functions tied to packet loss, retransmission rate, and bandwidth consumption. Adjustments are applied in real-time with fallback to rule-based logic if the learned policy violates operational thresholds. Unsupervised learning supports anomaly detection by clustering network telemetry data to identify deviations from normal patterns, enabling proactive identification of issues like connectivity drops or unusual traffic spikes. Additionally, unsupervised methods are used in virtual Bluetooth Low Energy (vBLE) location services; Juniper’s vBLE uses unsupervised Gaussian mixture models and density-based clustering (similar to DBSCAN) to resolve locations within 1–3 meters without requiring fixed beaconing hardware.

Cellular Networks: 5G Deployments and Pathways to 6G

5G networks incorporate ML/AI in operational settings for network optimization, traffic prediction, and fault detection. For instance, 3GPP Release 18 standards enable ML-based enhancements in radio access networks (RANs), including mobility optimization and resource slicing. These leverage supervised learning for predictive tasks, such as forecasting handover decisions based on historical mobility data, and unsupervised learning for clustering channel states to improve load balancing (3GPP TR 28.908). Deployed implementations demonstrate reduced handover failures through real-time adaptation to user patterns. In contrast, sixth-generation (6G) concepts remain primarily in research and early standardization phases. Initial stages, termed "AI for Network," explore ML for spectrum allocation and operations and maintenance (O&M) optimization, utilizing reinforcement learning (RL) for dynamic policy adjustments in simulated environments. Key ML techniques in deployed 5G systems include supervised learning for resource prediction and unsupervised learning for anomaly detection in modulation schemes. Reinforcement learning (RL) and federated learning (FL) are explored in research for dynamic environments, such as beamforming—where RL agents learn optimal beam selections via trial-and-error interactions with channel feedback—and privacy-preserving model training across distributed nodes. Generative AI, including large language models (LLMs), is investigated for semantic communication, where models compress data by extracting semantic features, but lacks widespread deployment.

NVIDIA Aerial for AI-Native Wireless R&D and Deployment

NVIDIA's Aerial platform supports both research and commercial deployments in 5G RANs. The CUDA-Accelerated RAN framework enables software-defined, GPU-accelerated networks, with full-inline acceleration of L1 and L2 layers. It incorporates ML techniques such as neural networks for beam management and channel state information (CSI) compression, where autoencoders reduce feedback overhead by learning compact representations of a channel data (see NVIDIA Aerial Documentation). The Omniverse Digital Twin (AODT) facilitates simulation for ML algorithm testing, supporting training of models like those for positioning via unsupervised clustering of signal patterns.

ML/AI in WAN And SD-WAN Predictive networks employ machine learning techniques to forecast network failures and SLA violations before they occur, contrasting with traditional reactive approaches that respond post-detection, based on Vasseur (2023). For failure prediction, models like Gradient Boosted Trees (GBT) incorporate features to identify patterns indicative of dark failures (complete connectivity loss) or grey failures (degraded performance). These algorithms are trained on large datasets from millions of paths across MPLS, Internet, DSL, fiber, satellite, and 4G links, optimizing for high precision to minimize “false positives”. Challenges involve balancing precision against recall ensuring robust alternate path testing to avoid oscillations. Deployed in hundreds of networks globally since 2023, these techniques improve Service Level Objectives (SLOs) and Quality of Experience (QoE), though specific model hyperparameters are not detailed in public sources. An example of predictive networking is implemented in Cisco's Catalyst SD-WAN WAN Insight feature, which applies similar ML/AI for analytics in software-defined wide area networks. Moreover, for bandwidth forecasting, statistical time-series forecasting models process historical usage data, including ingress and egress metrics aggregated daily, to predict future needs. These models incorporate seasonality and trend analysis, requiring weeks of historical data for generation, and output lower, upper, and mean bandwidth levels for comparison with actual usage (Cisco Catalyst SD-WAN Analytics Documentation).

ML/AI in Datacenters

NVIDIA's data center networking technologies apply ML/AI in several domains such as for congestion control, a power management, and predictive maintenance in AI workloads, based on public NVIDIA documentation and peer-reviewed research publications. For congestion control, reinforcement learning (RL) models—specifically deep RL with policy gradient optimization—learn optimal packet scheduling policies by processing network state inputs such as queue lengths, link utilization, and flow priorities to minimize latency and packet drops in RDMA over Converged Ethernet (RoCE) environments. These policies are distilled into decision tree representations to meet real-time performance constraints and are deployed efficiently on programmable network interface cards (NICs) such as ConnectX-6Dx. This distillation reduces inference latency by approximately 500×, enabling sub-2 µs decision times on NICs. Demonstrated in prototype clusters of up to 64 hosts operating at 100 Gbps with Spectrum-2 Ethernet switches, these RL-based congestion control systems outperform traditional approaches such as DCQCN and Swift in terms of latency, throughput, and fairness during AI training workloads (NVIDIA RL-CC, 2022; ML for Systems @ NeurIPS). For power management, supervised learning algorithms analyze telemetry data from GPUs and switches to predict and dynamically optimize power states. These models process metrics such as GPU load, temperature, and power draw to anticipate energy demands and adjust configurations accordingly. Exact model architectures are not publicly disclosed.

Optical Networks

Optical networking, encompassing dense wavelength-division multiplexing (DWDM) and coherent transceivers, presents opportunities for ML/AI in addressing nonlinear impairments and dynamic resource allocation. As of today, deployments remain limited, primarily in vendor-specific coherent optics for data center interconnects and metro networks, with most applications in research or pilot stages. Notable examples include Ciena's WaveLogic Ai and WaveLogic 6 platforms, which integrate ML for signal equalization and automation. Neural networks, such as convolutional architectures, model fiber nonlinearities by processing input features like signal-to-noise ratios and phase distortions, enabling adaptive compensation in deployed systems (Ciena WaveLogic Documentation). Ciena's implementations include neural network–based modules for OSNR estimation and nonlinear pre-distortion, deployed within the DSP pipeline of coherent modems, ennabling real-time compensation for impairments such as Kerr-induced nonlinear phase noise. These deployments demonstrate practical benefits in specific domains, yet they represent exceptions in broader networking practices.

3.2 Security: Deployed Applications of ML/AI in Threat Detection and Network Defense

For decades, network security has relied on foundational technologies like rule-based firewalls and Intrusion Detection Systems (IDS), which remain operational in many networks. These systems, exemplified by tools like Snort, rely on predefined signatures—essentially complex pattern-matching rules—to identify known threats in network traffic and block unauthorized access. However, this signature-based approach struggles with zero-day attacks and adaptive adversaries that use novel techniques to evade static patterns. This vulnerability is magnified by the emergence of generative AI, which enables adversaries to automate the creation of novel and polymorphic attacks, rendering signature-based detection increasingly ineffective. ML/AI techniques have been deployed at large scale to address these limitations, processing vast telemetry data for real-time threat detection and response. Vendors such as Cisco, Palo Alto Networks, and NETSCOUT (Arbor) integrate these methods into products handling security for millions of endpoints globally, augmenting traditional

systems with dynamic models trained on diverse datasets. For distributed denial-of-service (DDoS) attacks, unsupervised learning algorithms detect volumetric or application-layer floods by analyzing traffic anomalies. Density-based spatial clustering of applications with noise (DBSCAN) groups flow data based on features like packet rates, source IP diversity, and entropy measures. In Palo Alto's Prisma Cloud and NETSCOUT's Arbor Edge Defense, these models process netflow records in real time, incorporating hierarchical clustering to handle multi-dimensional data for mitigation actions like rate limitting (based on public documentation).

Data Exfiltration Detection use both unsupervised and supervised learning models like Random Forest classifiers to distinguish between legitimate and unauthorized outbound traffic. These models analyze features like destination ports and data volumes from network telemetry to identify and block potential data theft, as seen in products like Cisco Secure Network Analytics.

Ransomware Identification often employs techniques like Convolutional Neural Networks (CNNs) to analyze behavioral patterns from file activity and network traffic. By treating this data as image-like tensors, CNNs can detect malicious activities such as unauthorized encryption or command-and-control communication. This technique is used in platforms like Palo Alto's Cortex XDR to identify ransomware and its lateral movement.

Phishing detection leverages a wide variety of ML algorithms to parse email headers, URLs, and content for deceptive patterns. Deployed systems utilize a broad spectrum of techniques, including traditional models (e.g., Naive Bayes, Decision Trees, SVM, KNN), ensemble methods (Random Forests, Gradient Boosting), and deep learning architectures (CNNs, RNNs, Transformers). Hybrid models and Generative Adversarial Networks (GANs) are also employed to improve accuracy and resilience. For example, Cisco Secure Email applies these diverse techniques to inbound traffic for threat quarantine (Cisco Cybersecurity Report).

NVIDIA's security architecture leverages the BlueField DPU as a distributed, agentless sensor to enable a zero-trust posture. The Morpheus AI cybersecurity framework runs on the DPU, using GPU-accelerated inference to process raw packet telemetry at line rate (up to 100 Gbps) without host impact. This allows for AI-driven applications like digital fingerprinting and anomaly detection to be deployed directly in the network fabric, analyzing encrypted traffic streams in real time (NVIDIA Developer Blog).

Cisco's AI Endpoint Analytics uses ML/AI for detailed endpoint visibility and profiling. It employs supervised learning for multi-factor classification (MFC), using telemetry like DHCP fingerprints and DPI results to accurately label devices, including IoT. For unknown devices, it uses unsupervised clustering for "smart grouping," which creates profiling rules for policy enforcement via Cisco ISE. The platform also uses ML-based anomaly detection to identify spoofing by flagging behavioral deviations from established baselines. Detection of spoofing attack may trigger the device to be put under quarantine.

Device fingerprinting employs a wide range of ML algorithms—from traditional classifiers like Random Forests and SVMs to deep learning models like GNNs and autoencoders—to profile endpoints based on network behavior. These techniques are used, for example, by Cisco's AI Endpoint Analytics for IoT device identification. Closely related, User and Entity Behavior Analytics (UEBA) uses unsupervised models like Isolation Forests to baseline normal activity and detect

anomalies indicative of insider threats or compromised accounts. UEBA systems are often integrated into SIEM platforms to provide risk-scored alerts for prioritized investigation.

Cisco Encrypted Traffic Analysis (ETA) inspects encrypted traffic for threats without decryption by using ML models to analyze metadata. It analyzes features like the sequence of packet lengths and inter-arrival times (SPLT) and TLS handshake parameters. Recurrent Neural Networks (RNNs), particularly LSTM variants, are used to learn temporal patterns from this metadata to classify flows and detect threats like malware command-and-control channels while preserving privacy. The system can operate at up to 80 Gbps per sensor (Cisco Live EMEA, 2024).

3.3 Advanced Security Analytics and Automation

Beyond direct threat detection, ML/AI is also deployed to enhance security operations through predictive analytics and automation. Key applications include:

- **Vulnerability Prioritization:** NLP models like BERT analyze threat intelligence to predict which vulnerabilities are most likely to be exploited, helping teams prioritize remediation beyond static CVSS scores (e.g., Tenable, Qualys).
- **Security Orchestration, Automation, and Response (SOAR):** Platforms use topic modeling and machine learning classifiers to triage alerts and recommend response playbooks, automating SOC workflows (e.g., Palo Alto Networks' Cortex XSOAR).
- **Malware Propagation Tracking:** Graph Neural Networks (GNNs) model network connections to identify and track the spread of malware (e.g., Palo Alto's WildFire).
- **Insider Threat Monitoring:** Behavioral anomaly detection is used to identify deviations in user activity logs that could indicate a threat, flagging them for further investigation (e.g., Cisco Identity Services Engine).

These techniques have seen widespread deployment in enterprise and service provider environments, processing petabytes of daily traffic, but face challenges such as adversarial evasion through data poisoning and the need for continuous model retraining on imbalanced datasets. While the outlined algorithms enhance threat response, specific architectural hyperparameters and training protocols are not disclosed in public sources.

4. Untapped Potential of ML/AI in Networking

Despite a growing number of isolated deployments, the adoption of ML/AI in networking remains severely limited in both scope and depth. As documented in previous sections, current implementations are often confined to narrow, vendor-specific use cases within predefined operational boundaries. These examples represent valuable but shallow integration, falling far short of transforming the foundational mechanisms of how networks are operated, diagnosed, and optimized. For example, many critical functions—such as troubleshooting, root-cause-analysis, continuous optimization, and self-healing—have seen little to no fundamental innovation in decades, still relying on rule-based systems, static configurations, and manual intervention. The vision of autonomous / self-driving networks has been articulated in white papers and architectural roadmaps since at least the early 2000s, yet progress has been minimal, largely due to architectural inertia, data challenges, and risk aversion in production environments.

Modern networks generate massive volumes of telemetry that provide a rich data source for advanced ML/AI techniques to detect anomalies, model complex dependencies, and drive real-time control. Closing the gap between this potential and current deployment would unlock critical capabilities, including higher operational resilience, proactive root-cause diagnosis, predictive maintenance, QoE-aware routing, and cross-layer optimization. Below, we outline several categories of high-impact use cases where ML/AI could play a transformative role.

Optimization Use Cases Traffic Engineering: ML models could predict congestion and dynamically adjust routing paths. In inter-domain routing, reinforcement learning agents could replace brittle manual techniques—like AS-PATH prepending and MED tuning—with adaptive policies learned from continuous feedback. Unlike traditional BGP-based traffic engineering, which depends on statically defined heuristics, RL-based systems could adjust policies in real time, optimizing for latency, packet loss, or throughput across ever-changing topologies.

Spectrum Allocation in Wireless Networks: Machine learning enables cognitive radio systems to identify underutilized spectrum in real time, improving spectral efficiency in dense 5G and 6G environments. Cisco's AI-Enhanced Radio Resource Management (RRM), for example, integrates cloud-based ML to optimize transmit power and channel assignment based on long-term telemetry, surpassing traditional reactive, heuristic-based RF tuning.

Fault Management and Reliability Use Cases Predictive Maintenance: various ML/AI technologies can detect early warning signals of hardware degradation or link/node/path failures. Models trained on historical telemetry can issue maintenance alerts before service degradation occurs.

Root-Cause Analysis: Current approaches to Root-Cause Analysis (RCA) are often limited to basic correlation of events, but it is well known that correlation does not imply causation. There is a massive untapped potential for more sophisticated root-causing by using a combination of classic ML/AI and Generative AI. These advanced systems could leverage external tools, analyze complex dependencies, and move beyond simple correlation to identify the true origin of network issues.

Self-Healing Networks: AI agents embedded in SDN controllers or network fabric elements could detect, localize, and respond to faults in real time by rerouting traffic, restarting subsystems, or reallocating resources—without human intervention. While the concept has been discussed for decades, practical implementations are rare, highlighting a critical area of unmet potential.

Cross-Domain Optimization: Most production environments still treat compute, storage, and network as independent silos. ML models trained on telemetry from all these domains could provide a holistic understanding of system-wide dependencies, enabling a far more accurate root-cause analysis that is simply not possible with today's siloed monitoring tools.

Cross-Layer Optimization: ML/AI can be used to understand the implications of an anomaly at one layer of the network stack on another. For example, a model could learn the relationship between network jitter at the transport layer and application response time at Layer 7. A more complex use case would be to assess the impact of packet loss on the performance of a distributed AI training job, allowing for proactive adjustments to the network fabric or the workload itself. Such insights are nearly impossible to codify with traditional rule-based systems.

These are only a few examples of the potential applications of ML/AI in networking.

Moreover, the most significant catalyst for unlocking the full potential of ML/AI in networking is the recent emergence of Generative AI and agentic architectures, which represent a paradigm shift we will explore in detail later in this document.

5. Challenges and Barriers to ML/AI Adoption in Networking

To provide a balanced view, it is essential to address the practical obstacles that have hindered wider adoption of ML/AI in networking. This section discusses key challenges, drawing from operational realities in network environments, and notes potential paths forward. Understanding these barriers is crucial for realizing the potential outlined earlier.

Cultural and Operational Resistance: A significant, though often understated, barrier is the cultural and operational resistance from seasoned network engineers. Networking has traditionally been a discipline rooted in deterministic, command-line-driven configurations and explicit rule-sets where predictability is paramount. The introduction of ML/AI, with its probabilistic nature and "black-box" models, can be perceived as a loss of control and a threat to the stability that engineers have spent careers ensuring. This skepticism is compounded by concerns about job roles evolving or becoming obsolete, as well as a deep-seated professional ethos that prioritizes manual intervention and deep, protocol-level understanding over automated, data-driven decision-making, especially in high-stakes environments where a single error can cause widespread outages.

Data Quality and Availability: Network-generated data often suffers from incompleteness, noise, or bias due to varying device capabilities and intermittent connectivity. For example, telemetry from edge devices may lack standardization, leading to models that perform poorly in diverse settings.

Integration with Legacy Infrastructure: Many existing networks use proprietary or outdated protocols, making it difficult to integrate modern ML/AI frameworks without significant and risky retrofitting. Deploying AI agents in such environments can introduce compatibility issues and performance overhead. Addressing these challenges through standardized data formats and hybrid systems is pivotal for transitioning to more intelligent networks.

6. On The Emergence of Generative AI and Large Language Models The adoption of large language models (LLMs) and generative AI (Gen-AI) is transforming the landscape of modern computing. Their integration across all fields such as software development, business automation, healthcare, manufacturing and scientific discovery is driven by an ecosystem of rapidly evolving techniques: Retrieval-Augmented Generation (RAG) augments LLMs with external knowledge sources, improving factual grounding and contextual relevance. Advanced designs like Graph-RAG, Multi-hop RAG, and MAIN-RAG allow multi-step reasoning and structural retrieval, enabling deeper question answering and decision support. Chain-of-Thought (CoT), Tree-of-Thought, and Graph-of-Thought Reasoning introduce explicit intermediate reasoning steps. These frameworks help LLMs break down complex problems into logical sequences, allowing interpretable and robust decision-making across planning, diagnostics, and inference. Mixture-of-Experts (MoE) architectures activate sparse subsets of expert subnetworks during inference, scaling model capacity while maintaining inference efficiency. This makes it possible to support diverse tasks while minimizing compute cost. Test-Time Adaptive Computation techniques, including early exits, speculative decoding, and token-

level routing, enable models to dynamically allocate resources based on input complexity. This improves latency, energy efficiency, and responsiveness, particularly in edge and mobile environments. Short-Term Memory and Context Management techniques—such as scratchpad prompting, memory tokens, context window extension, and compressive attention—support continuity and coherence in long multi-turn tasks. These mechanisms allow models to reason over extended sequences of actions and events. Multi-Turn and Long-Horizon Planning has been enhanced through agent frameworks like LangChain, AutoGen, and LangGraph. These tools support memory persistence, dynamic state tracking, and autonomous goal decomposition across evolving task contexts. Reinforcement Learning (RL) has evolved beyond static reward shaping (as used in RLHF).

Agentic Systems represent a new paradigm where LLM-powered agents possess memory, planning, and collaboration capabilities. While poised to have a drastic impact, this is not "automagic"; realizing this potential will require careful design to address the new challenges these systems introduce. These agents can:

- Coordinate with one another to solve distributed tasks
- Learn to use external tools (e.g., APIs, search engines, code interpreters) through emergent skill acquisition
- Access modular reasoning capabilities by combining symbolic logic, search, and probabilistic models
- Engage in self-improvement, using methods like reflective prompting, automatic chain-of-thought revision, and online fine-tuning based on feedback and outcomes

Small LLMs (e.g., 1B–7B parameters) are increasingly deployed at the edge. Using quantization, pruning, distillation, and LoRA fine-tuning, these models perform local inference on switches, mobile devices, and embedded platforms, enabling intelligence outside cloud boundaries. Together, these advances make Gen-AI not only more powerful, but more adaptable, distributed, and autonomous—setting the stage for a fundamental transformation in how intelligent systems interact with the world.

Implications for Networking while current networking applications of LLMs remain limited—mostly in assistive roles such as configuration generation, documentation parsing, or chatbot-driven support—the trajectory of Gen-AI points toward much deeper integration:

- **Embedded intelligence:** Compact LLMs or agentic components could reside inside GPU/CPU, routers, switches, or virtualized infrastructure, enabling real-time inference and autonomous control at the edge.
- **Root-cause analysis and debugging:** LLMs equipped with RAG and reasoning could correlate logs, telemetry, and alerts across protocol layers to diagnose complex network events without human intervention.
- **Cross-layer reasoning:** Traditional networking architectures treat OSI layers as separate domains. Gen-AI models could reason across transport, application, and physical layers to detect interactions, performance bottlenecks, or emergent failure modes.
- **Distributed agent collaboration:** Agentic systems could coordinate routing decisions, optimize policy configurations, and dynamically reallocate resources by communicating across administrative domains, achieving self-organizing behavior.

This transition—away from static, centralized control planes toward intelligent, distributed, agent-driven networks—will mark a major inflection point in in Internet architecture. Intelligence will become

embedded in the fabric of the network, learning from experience, adapting over time, and collaborating across nodes and layers. The impact of these trends goes beyond incremental enhancement.

We are entering an era in which the foundational design of the Internet itself may be reimagined: from a passive data transport layer to an active, cognitive, and self-optimizing infrastructure. These architectural implications—including protocol design, trust models, AI-native control planes, and agent security—will be addressed in a forthcoming white paper dedicated to the AI-driven reinvention of the Internet.

7. Conclusion: From Local Optimizations to Smarter Networks

ML/AI have already delivered very valuable results in networking, but only in a few well-defined areas. For instance, ML is used to detect anomalies in Wifi networks, allow for predictive actions in SD-WAN, optimize radio frequency parameters in Wi-Fi networks, detect anomalies in telemetry data, and predict links/nodes/pathsfailures.

In network security, an area that has arguably seen the broadest adoption of ML/AI, techniques have been successfully applied to identify DDoS attacks using clustering algorithms, detect ransomware with neural networks, and flag insider threats through behavioral analysis. However, even here, these applications often remain vendor-specific and have yet to realize the full potential of truly intelligent, adaptive networks.

The real opportunity lies ahead. Generative AI and large language models (LLMs) bring a new kind of capability. They make it possible to design distributed, intelligent systems that adapt in real time, reason across multiple layers, and take decisions proactively. Instead of relying on static rules and manual processes, networks could become self-aware and self-optimizing, reacting to problems before users even notice them.

However, a new gap is emerging: the gap between what GenAI can offer and what has actually been adopted in networking is even wider. While other industries—like software development, finance, healthcare, and scientific computing—are rapidly evolving with tool-using agents, reasoning models, and advanced planning algorithms, networking use remains mostly assistive. Today's GenAI applications in networking are largely limited to configuration help, documentation search, or simple chatbot interfaces. The potential to deploy LLMs for autonomous troubleshooting, real-time root cause analysis, or distributed policy negotiation is still largely unrealized. The pressure to evolve is growing. With rising data volumes, increasing complexity, and higher reliability demands, legacy systems are becoming harder to maintain and scale. Without a shift toward intelligent, learning-based infrastructures, networking risks falling behind the pace of transformation happening across the digital ecosystem. Now is the time to close the gap between what networks are—and what they could become.

The views and opinions expressed in this paper are solely those of the author and do not necessarily reflect the official policy or position of NVIDIA or any of its affiliates. The information and analysis presented are based exclusively on publicly available data, documentation, and reports accessible on the web.

8. Appendix: Selected References

For further reading on concepts discussed in this paper, consider the following resources:

Vendor-Specific AI/ML Documentation:

- **Cisco Systems:**
  - Cisco AI Network Analytics Overview. (Details on baselining KPIs and anomaly detection within Cisco Catalyst Center). Available at: https://www.cisco.com/c/en/us/products/dna-analytics-and-assurance.html
  - ThousandEyes WAN Insights. (Covers the use of predictive analytics for SD-WAN path optimization). Available at: https://www.thousandeyes.com/product/wan-insights
  - Cisco AI Endpoint Analytics White Paper. (Details the use of ML for device profiling and security). Available at: https://www.cisco.com/c/en/us/solutions/collateral/enterprise-networks/software-defined-access/nb-06-ai-endpoint-analytics-wp-cte-en.html
  - Cisco Secure Network Analytics (Stealthwatch) At-a-Glance. (Covers ML for detecting threats in encrypted traffic). Available at: https://www.cisco.com/c/en/us/products/collateral/security/stealthwatch/secure-network-analytics-aag.html
- **Juniper Networks / Mist Systems:**
  - The AI-Driven Campus Architecture White Paper. (Explains the application of Mist AI for wired and wireless assurance and the role of the Marvis VNA). Available at: https://www.juniper.net/content/dam/www/assets/white-papers/us/en/the-ai-driven-campus-architecture.pdf
  - Get Started with Marvis Documentation. (Official documentation for the Marvis conversational assistant and its troubleshooting capabilities). Available at: https://www.juniper.net/documentation/us/en/software/mist/mist-aiops/topics/concept/marvis-actions-overview.html
- **NVIDIA:**
  - NVIDIA Morpheus Developer Page. (Provides details on the AI cybersecurity framework). Available at: https://developer.nvidia.com/morpheus
  - NVIDIA DOCA SDK Developer Page. (Official resource for the DPU programming framework). Available at: https://developer.nvidia.com/networking/doca/overview
- **Palo Alto Networks:**
  - Cortex XDR Product Page. (Describes the use of ML for endpoint and network threat detection). Available at: https://docs-cortex.paloaltonetworks.com/p/XDR
- **NETSCOUT / Arbor:**
  - Arbor Edge Defense (AED) Datasheet. (Details the use of ML for automated DDoS threat detection). Available at: https://www.netscout.com/sites/default/files/2024/02/data-sheet-arbor-edge-defense-1606-0622.pdf
- **Ciena:**
  - WaveLogic Ai Product Page. (Provides details on the programmable coherent optics and the real-time link monitoring data it enables). Available at: https://www.ciena.com/products/wavelogic/wavelogic-ai
  - WaveLogic 6 Technology Overview. (Describes the next generation of coherent optics, focusing on performance, scalability, and efficiency for 800G and 1.6T). Available at: https://www.ciena.com/products/wavelogic/wavelogic-6